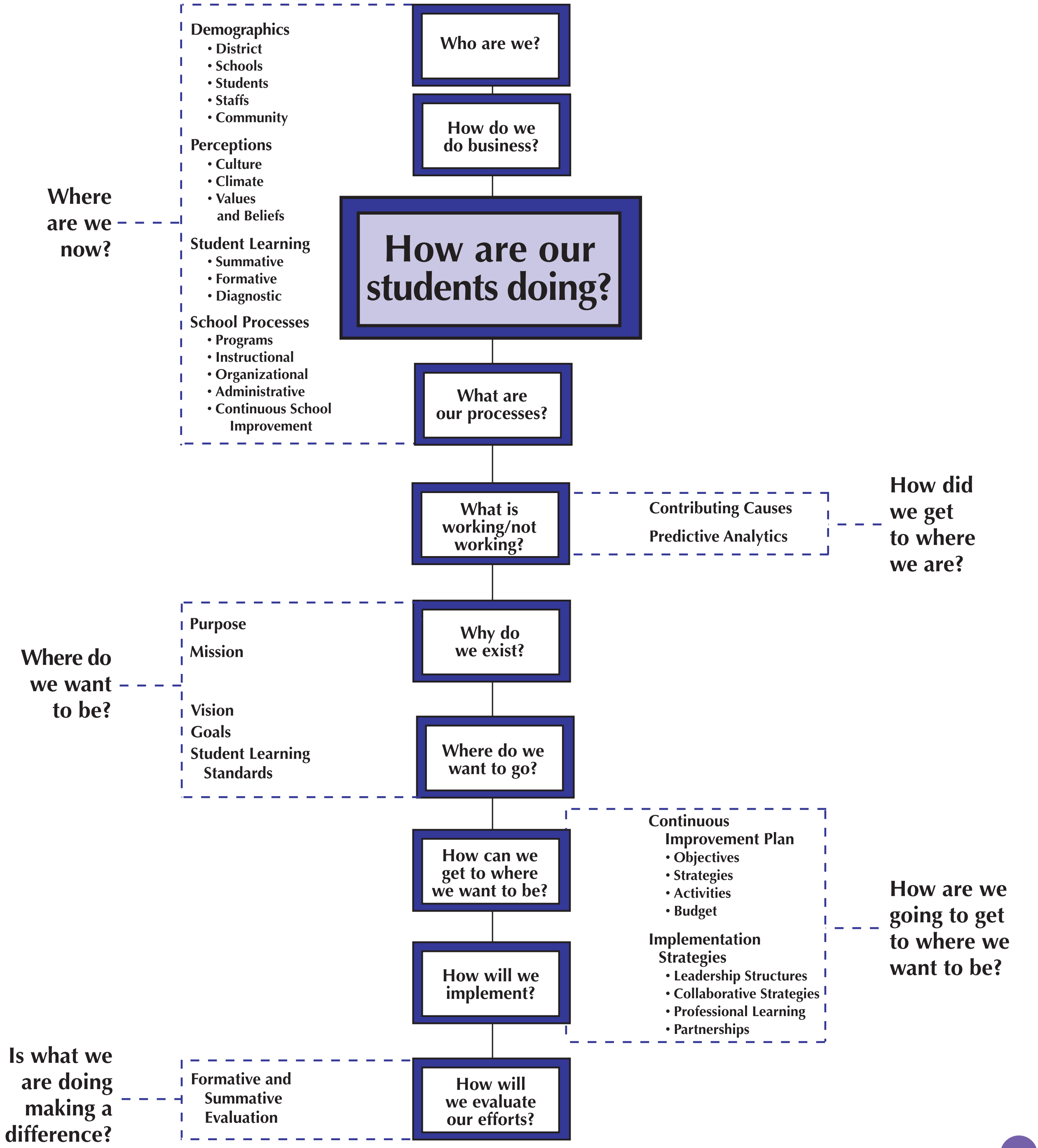
HOW ARE OUR STUDENTS DOING: STUDENT LEARNING DATA

5

CHAPTER



Learning does not take place in isolation. Students bring to the learning setting what they have experienced and the values they have been taught at home and in their neighborhoods. This effects how they respond.

National Center for Education Statistics

When schools focus on compliance, they typically use only summative measures to judge their progress. Schools committed to continuous school improvement use multiple and ongoing measures of data to understand what students know as a result of instruction, what teachers are teaching, and which students need extra help and on what. These schools use student learning measures to provide valuable information for adjusting instruction to meet the needs of *all* students.

> ***Schools committed to continuous school improvement use multiple and ongoing measures of data to understand what students know as a result of instruction, what teachers are teaching, and which students need extra help and on what.***

STUDENT LEARNING DATA: WHAT THEY ARE AND WHY THEY ARE IMPORTANT TO CONTINUOUS SCHOOL IMPROVEMENT

How are our students doing?, a sub-question within the continuous school improvement framework question, *Where are we now?,* requires a synthesis of student learning data in all subject areas, disaggregated by student demographic groups, by teachers, by grade levels, by following the same groups of students (cohorts) over time, as well as looking at individual student growth. Student learning data show if schools are meeting the needs of all students and uncover strengths in learning and areas for improvement. If students are not proficient, teachers need to know what the students know and what they do not know. If students are not proficient, teachers need to know how many are not proficient, and by how much students must improve to become proficient. Looking at student learning data within grade levels, and looking at student learning across grade levels show if a school has instructional coherence, the alignment of curriculum, instruction, and assessment within grade levels, and across grade levels to create a continuum of learning for students.

> ***Student learning data show if schools are meeting the needs of all students and uncover strengths in learning and areas for improvement.***

MEASURES OF STUDENT LEARNING

Assessment plays a major role in how students learn, their motivation to learn, and how teachers teach. Assessment is used for three main purposes, as succinctly described in this Manitoba Education website: *http://www.edu.gov.mb.ca/k12/assess/wncp/rethinking_assess_mb.pdf.*

The Role of Assessment in Learning

- ***Assessment* FOR *learning:*** where assessment helps teachers gain insight into what students understand in order to plan and guide instruction, and provide helpful feedback to students.
- ***Assessment* AS *learning:*** where students develop an awareness of how they learn and use that awareness to adjust and advance their learning, taking an increased responsibility for their learning. (*Note:* This is essential for successful Common Core State Standards implementation.)
- ***Assessment* OF *learning:*** where assessment informs students, teachers and parents, as well as the broader educational community, of achievement at a certain point in time in order to celebrate success, plan interventions, and support continued progress.

Assessment must be planned with its purpose in mind. Assessment **FOR**, **AS**, and **OF** learning all have a role to play in supporting and improving student learning, and must be appropriately balanced. The most important part of assessment is the interpretation and use of the information that is gleaned for its intended purpose.

Assessment is embedded in the learning process. It is tightly interconnected with curriculum and instruction. As teachers and students work toward the achievement of curriculum outcomes, assessment plays a constant role in informing instruction, guiding the student's next steps, and checking progress and achievement. Teachers use many different processes and strategies for classroom assessment, and adapt them to suit the assessment purpose and needs of individual students.

Research and experience show that student learning is best supported when—

- Instruction and assessment are based on clear learning goals.
- Instruction and assessment strategies are differentiated according to student learning needs.
- Students are involved in the learning process (they understand the learning goal and the criteria for quality work, receive and use descriptive feedback, and take steps to adjust their performance) and learn to self-regulate.
- Assessment information is used to make decisions that support further learning.
- Parents are well informed about their child's learning, and work with the school to help plan and provide support.
- Students, families, and the general public have confidence in the system.

Reproduced from Manitoba Education. "The Role of Assessment in Learning." *Assessment and Evaluation. www.edu.gov.mb.ca/k12/assess/role.html* (January 2013).

WAYS TO MEASURE STUDENT LEARNING

Comparing results on different measures gives teachers insight into what teaching strategies, as well as testing strategies, work best with different students. Teachers can also determine which formative assessments best prepare students for summative tests. Appendix B3, the *Student Learning Data Inventory*, helps schools organize their assessments within and across grade levels and subject areas, by assessment type. Typical assessment terms, organized by assessments for, as, and of learning, are defined in Figure 5.1.

Comparing results on different measures gives teachers insight into what teaching strategies, as well as testing strategies, work best with different students.

Figure 5.1
ASSESSMENT DEFINITIONS AND USES

Assessments* FOR *Learning are in-class measures, usually given frequently, to determine student success with the curriculum and for teachers to determine what additional or modified instruction is needed. (Stiggins, R. J. (1999). Teams. *Journal of Staff Development,* 20(3), 17-21).

Definitions	Cautions
Classroom assessment. An assessment developed, administered, and scored by a teacher or set of teachers with the purpose of evaluating individual or classroom student performance on a topic. Classroom assessments may be aligned into an assessment system that includes alternative assessments and norm-referenced and/or criterion-referenced assessments. (Center for Research on Evaluation, Standards, and Student Testing [CRESST], retrieved 08/26/12)	Ideally, the results of a classroom assessment are used to inform and influence instruction that helps students reach high standards.
Diagnostic tests, usually standardized and normed, are given before instruction begins to help the instructor(s) understand student learning needs.	Diagnostic tests can help teachers know the nature of students' difficulties, through sub-domain performance. Many different score and question types are used with diagnostic tests.
Formative assessment is a range of formal and informal assessment procedures employed by teachers during the learning process in order to modify teaching and learning activities to improve student attainment.	When incorporated into classroom practice, formative assessments provide the information needed to adjust teaching and learning while they are happening. In this sense, formative assessments inform both teachers and students about student understanding at a point when timely adjustments can be made. These adjustments help to ensure students achieve targeted, standards-based learning goals within a set time frame.
Progress monitoring assessments are used to assess students' academic performance, to quantify a student rate of improvement or responsiveness to instruction, and to evaluate the effectiveness of instruction. Progress monitoring can be implemented with individual students or an entire class. (Retrieved from www.rti4success.org/pdf/glossry_of_terms, 11/06/12)	Progress monitoring assessments allow a teacher to track students in a specific skill area, or they could be more general tests of grade level curricula. Progress monitoring is conducted at least monthly, but often more frequently, to (a) estimate rates of improvement, (b) identify students who are not demonstrating adequate progress, and/or (c) compare the efficacy of different forms of instruction to design more effective, individualized instruction.
Screeners involve brief assessments that are valid, reliable, and evidence-based. They are typically applied three times per year.	Screeners are conducted with all students or targeted groups of students to identify students who are at risk of academic failure and, therefore, likely to need additional or alternative forms of instruction to supplement the conventional general education approach.

Figure 5.1 *(Continued)*
ASSESSMENT DEFINITIONS AND USES

Assessments* AS *Learning are in-class measures, usually given frequently, to determine student success with the curriculum and for teachers to determine what additional or modified instruction is needed. (Stiggins, R. J. (1999). Teams. *Journal of Staff Development*, 20(3), 17-21).

Definitions	Cautions
Alternative assessments (**AKA authentic assessment, performance assessment**) are assessments that require students to generate a response to a question rather than choose from a set of responses provided to them. Exhibitions, investigations, demonstrations, written or oral responses, journals, and portfolios are examples of the assessment alternatives we think of when we use the term "alternative assessment." Alternative assessments are usually one element of an assessment system. (CRESST, retrieved 08/26/12)	Ideally, alternative assessments require students to actively accomplish complex and significant tasks, while bringing to bear prior knowledge, recent learning, and relevant skills to solve realistic or authentic problems.
Performance assessments refer to assessments that measure skills, knowledge, and ability directly-such as through performance. In other words, if you want students to learn to write, you assess their ability on a writing activity.	One must find a way to score these results and make sense for individual students and groups of students.
A **rubric** is a scoring tool that judges work against a set of criteria. A rubric divides, on a descriptive scale, the assigned work into component parts and provides clear descriptions of the characteristics of the work associated with each component, at varying levels of mastery.	Rubrics can be used for a wide array of assignments: papers, projects, oral presentations, artistic performances, group projects, etc. Rubrics can be used as scoring or grading guides, to provide formative feedback to support and guide ongoing learning efforts, or both.
Standardized tests are assessments that have uniformity in content, administration, and scoring.	Standardized tests can be used for comparing results across students, classrooms, schools, school districts, and states. Norm-referenced, criterion-referenced, and diagnostic tests are the most commonly used standardized tests. Psychometric and edumetric standardized tests are often used to identify children with special needs.

Figure 5.1 *(Continued)*
ASSESSMENT DEFINITIONS AND USES

Assessments* OF *Learning inform students, teachers, and parents, as well as the broader educational community, of achievement at a certain point in time in order to celebrate success, plan interventions, and support continued progress.

Definitions	Cautions
Benchmark. A detailed description of a specific level of performance expected of students at particular ages, grades, or development levels. Benchmarks are often represented by scoring rubrics and exemplars of student work. (CRESST, retrieved 08/26/12)	A set of benchmarks can be used as "checkpoints" to monitor progress toward meeting performance goals within and across grade levels, i.e., benchmarks for expected mathematics capabilities at Grades 3, 7, 10, graduation.
Criterion-referenced measures compare an individual's performance to a specific learning objective or performance standard and not to the performance of other test takers. Criterion-referenced assessments tell us how well students are performing on specific criteria, goals, or standards. (CRESST, retrieved 08/26/12)	For school level analyses, criterion-referenced tests are usually scored on descriptive scales in terms of the number or percentage of students meeting the standard or criterion, or the number or percentage of students falling in typical descriptive categories, such as far below basic, below basic, basic, proficient, and advanced. Criterion-referenced tests can be standardized or not, and they can also have norming groups.
End-of-Course Exams, Including Certification Exams End-of-Course Exams are criterion-referenced tests given at the completion of a course of study.	End-of-course exams are given to determine whether a student demonstrates attainment of the knowledge and skills necessary for mastery of that subject.
Norm-referenced tests are standardized tests. Norm-referenced test scores create meaning through comparing the test performance of a school, group, or individual, with the performance of a norming group. A norming group is a representative group of students whose results on a norm-referenced test help create the scoring scales with which others compare their performance. Norming groups' results are professed to look like the normal curve, shown in Figure 5.2.	Norm-referenced test results are used to understand how your students scored in comparison to the norming group.
Outcome assessments are given at the end of the school year. Outcome tests are frequently group-administered tests of important outcomes.	Outcome assessments are often used for district, state, and provincial reporting purposes. These tests are important because they give school leaders and teachers feedback about the overall effectiveness of their instructional program. As part of an effective assessment plan, outcome assessments should be administered at the end of every year.
Provincial assessments are mostly outcome assessments that vary by province.	Provincial assessments are used to tell schools and boards how students are performing with respect to specific subjects.
Standards-based assessments assess student achievement on the basis of outcomes or standards performance. They are usually criterion-referenced.	Standards-based assessments use cut scores to know if standards have been met. Most common uses are to know the number or percentage of students meeting or exceeding a standard.

Figure 5.1 *(Continued)*
ASSESSMENT DEFINITIONS AND USES

Definitions	Cautions
Standards-based assessments assess student achievement on the basis of outcomes or standards performance. They are usually criterion-referenced.	Standards-based assessments use cut scores to know if standards have been met. Most common uses are to know the number or percentage of students meeting or exceeding a standard.
Standardized tests are assessments that have uniformity in content, administration, and scoring.	Standardized tests can be used for comparing results across students, classrooms, schools, school districts, and states. Norm-referenced, criterion-referenced, and diagnostic tests are the most commonly used standardized tests. Psychometric and edumetric standardized tests are often used to identify children with special needs.
Summative assessments provide a bottom line of learning as related to established grade-level standards and norms. Most summative assessments are given yearly or pre-and post during a year.	Summative assessments are developed and scored in ways that ensure reliability and validity, and may be norm or criterion referenced. Summative or outcome assessments are used for student screening, accountability, or pre-post measures of the efficacy of programs.
Teacher-assigned grades. Teachers use number or letter grades to judge the quality of a student's performance on a task, unit, or during a period of time. Grades are most often given as A, B, C, D, F, (or R, 1, 2, 3, 4), with pluses and minuses given by some teachers for the first four to distinguish among students.	Grades are mostly used to tell students and parents how well the students did on a task, or during a period of time. Grades mean different things to different teachers. Needless to say, grades can be subjective.

Student learning data come from screening assessments, diagnostic assessments, classroom assessments, classroom assignments and activities, formative assessments, state/provincial assessments, performance and standards assessments, and grades. In a perfect scenario, teachers are clear and agree on what they want students to know and be able to do by the end of the year, course, or lesson, also referred to as student learning standards. Short screening assessments, or screeners, administered to all students help teachers know if students are at risk of failure. For students who did not perform well on the screeners, teachers use longer diagnostic assessments to understand what the students do and do not know, and consider how they can help each student with her/his learning needs. Teachers plan for instruction for all students with this information. On a regular basis, teachers assess to understand what students are learning and which students need extra support, and adjust instruction to meet all student needs. Teachers then assess to know if the students learned what they were expected to learn.

Student learning data come from screening assessments, diagnostic assessments, classroom assessments, classroom assignments and activities, formative assessments, state/provincial assessments, performance and standards assessments, and grades.

Analyses of all types of student learning measures used in the school can help one know if all students are learning, and if true learning can be detected better with one measure than another.

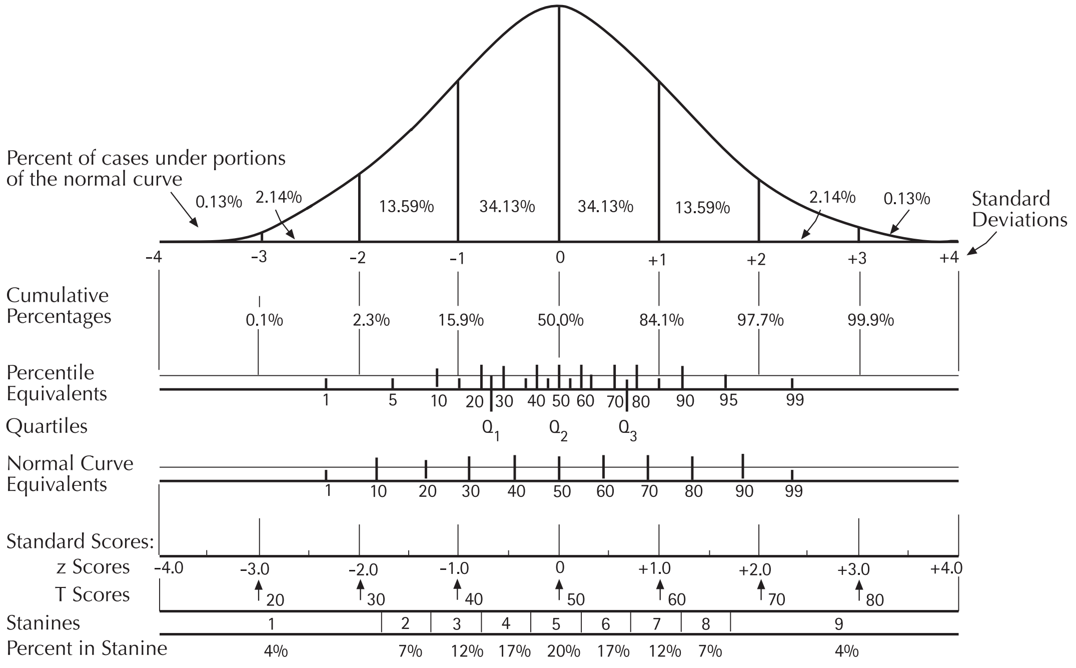
Analyses of all types of student learning measures used in the school can help one know if all students are learning, and if true learning can be detected better with one measure than another. Looking across measures, teachers can determine how the different measures compare in performance and if students perform differently on one type of test versus another.

THE NORMAL CURVE AND RELATED SCORE TYPES

In a normal distribution, the measures are distributed symmetrically about the mean, or average score.

A text on data analysis would not be complete if it did not include a section on the normal curve, and the scores associated with it. The normal curve (Figure 5.2) is a distribution of scores or other measures that in graphic form has a distinctive bell-shaped appearance. In a normal distribution, the measures are distributed symmetrically about the mean, or average score. Most results are near the mean and decrease in frequency the farther one departs from the mean. Stated another way—the theory of the normal curve basically says that when a test is given to a representative sample of students, most students score around the mean with few students scoring very high or very low. Using this theory, test publishers are able to create scales that are useful for schools to compare their scores with the norming group.

Figure 5.2
THE NORMAL CURVE AND SCORES



Stanford Achievement Test: Eighth Edition. Copyright © 1989 by
The Psychological Corporation. Reproduced by permission. All rights reserved.

National Percentile Ranks (NPR). The two most commonly used and useful normed scales, or score types, are national percentile ranks and normal curve equivalents. The national percentile rank, also known as the national percentile equivalent, ranges from 1 to 99, with a midscore of 50. The NPR is one of the most used scales; it is also misused the most. The misuse of this scale stems from the fact that it is an unequal interval scale, which prohibits adding, subtracting, multiplying, and dividing the scores. *One should not look at gains, losses, or averages with percentile ranks* because of the unequal interval scale, and because the scores are dependent on how the norming group performed. If a student performed at the 70th percentile last year and 75th percentile this year—that is how the performance is described. You can see on the normal curve (Figure 5.2) how her/his performance has changed with respect to the norming group. Median scores, or the middle scores, are the most appropriate means of describing a whole school's typical performance.

The two most commonly used and useful normed scales, or score types, are national percentile ranks and normal curve equivalents.

The national percentile rank, also known as the national percentile equivalent, ranges from 1 to 99, with a midscore of 50.

Normal Curve Equivalent Scores (NCE). Normal curve equivalent (NCE) scores were created by educational researchers to alleviate the problem of unequal interval scales. This equal interval scale has a mean of 50, and range of 1 to 99, just like the NPR. A standard deviation of 21.06 is used to ensure that NCE and percentile ranks have equivalent scores at the 1st, 50th, and 99th percentiles. NCEs have the same meaning across students, subtests, grade levels, classrooms, schools, and school districts. Fifty (50) is what one would expect for an average year's growth; put another way, 50 is grade level—always the national average for the grade and month the test is taken. You can look at how close your scores are to expected performance, averages, gains, losses, highest, and lowest scores. NCEs are excellent for looking at scores over time.

Normal curve equivalent (NCE) scores were created by educational researchers to alleviate the problem of unequal interval scales.

Convert percentile ranks to NCEs so you can compare scores over time, or NCEs to percentiles to understand how your students did in relationship to other similar students, using the tables in Figures 5.3 and 5.4. Note how the conversion can be more exact going from NCE scores to percentiles. A conversion in the opposite direction uses an average NCE.

Figure 5.3
NCE TO PERCENTILE CONVERSION TABLE

NCE Range	Percentile Rank	NCE Range	Percentile Rank	NCE Range	Percentile Rank	NCE Range	Percentile Rank
1.0 – 4.0	1	36.1–36.7	26	50.3–50.7	51	64.6–65.1	76
4.1 – 8.5	2	36.8–37.3	27	50.8–51.2	52	65.2–65.8	77
8.6–11.7	3	37.4–38.0	28	51.3–51.8	53	65.9–66.5	78
11.8–14.1	4	38.1–38.6	29	51.9–52.3	54	66.6–67.3	79
14.2–16.2	5	38.7–39.2	30	52.4–52.8	55	67.4–68.0	80
16.3–18.0	6	39.3–39.8	31	52.9–53.4	56	68.1–68.6	81
18.1–19.6	7	39.9–40.4	32	53.5–53.9	57	68.7–69.6	82
19.7–21.0	8	40.5–40.9	33	54.0–54.4	58	69.7–70.4	83
21.1–22.3	9	41.0–41.5	34	54.5–55.0	59	70.5–71.3	84
22.4–23.5	10	41.6–42.1	35	55.1–55.5	60	71.4–72.2	85
23.6–24.6	11	42.2–42.7	36	55.6–56.1	61	72.3–73.1	86
24.7–25.7	12	42.8–43.2	37	56.2–56.6	62	73.2–74.1	87
25.8–26.7	13	43.3–43.8	38	56.7–57.2	63	74.2–75.2	88
26.8–27.6	14	43.9–44.3	39	57.3–57.8	64	75.3–76.3	89
27.7–28.5	15	44.4–44.9	40	57.9–58.3	65	76.4–77.5	90
28.6–29.4	16	45.0–45.4	41	58.4–58.9	66	77.6–78.8	91
29.5–30.2	17	45.5–45.9	42	59.0–59.5	67	78.9–80.2	92
30.3–31.0	18	46.0–46.5	43	59.6–60.1	68	80.3–81.7	93
31.1–31.8	19	46.6–47.0	44	60.2–60.7	69	81.8–83.5	94
31.9–32.6	20	47.1–47.5	45	60.8–61.3	70	83.6–85.5	95
32.7–33.3	21	47.6–48.1	46	61.4–61.9	71	85.6–88.0	96
33.4–34.0	22	48.2–48.6	47	62.0–62.5	72	88.1–91.0	97
34.1–34.7	23	48.7–49.1	48	62.6–63.1	73	91.1–96.4	98
34.8–35.4	24	49.2–49.6	49	63.2–63.8	74	96.5–99.0	99
35.5–36.0	25	49.7–50.2	50	63.9–64.5	75		

Stanford Achievement Test: Eighth Edition. Copyright © 1989 by The Psychological Corporation. Reproduced by permission. All rights reserved.

Figure 5.4
PERCENTILE TO NCE CONVERSION TABLE

Percentile Rank	NCE	Percentile Rank	NCE	Percentile Rank	NCE	Percentile Rank	NCE
1	1.0	26	36.5	51	50.5	76	64.9
2	6.7	27	37.1	52	51.1	77	65.6
3	10.4	28	37.7	53	51.6	78	66.3
4	13.1	29	38.3	54	52.1	79	67.0
5	15.4	30	39.0	55	52.6	80	67.7
6	17.3	31	39.6	56	53.2	81	68.5
7	18.9	32	40.1	57	53.7	82	69.3
8	20.4	33	40.7	58	54.2	83	70.1
9	21.8	34	41.3	59	54.8	84	70.9
10	23.0	35	41.9	60	55.3	85	71.8
11	24.2	36	42.5	61	55.9	86	72.8
12	25.3	37	43.0	62	56.4	87	73.7
13	26.3	38	43.6	63	57.0	88	74.7
14	27.2	39	44.1	64	57.5	89	75.8
15	28.2	40	44.7	65	58.1	90	77.0
16	29.1	41	45.2	66	58.7	91	78.2
17	29.9	42	45.8	67	59.3	92	79.6
18	30.7	43	46.3	68	59.9	93	81.1
19	31.5	44	46.8	69	60.4	94	82.7
20	32.3	45	47.4	70	61.0	95	84.6
21	33.0	46	47.9	71	61.7	96	86.9
22	33.7	47	48.4	72	62.3	97	89.6
23	34.4	48	48.9	73	62.9	98	93.3
24	35.1	49	49.5	74	63.5	99	99.0
25	35.8	50	50.0	75	64.2		

Stanford Achievement Test: Eighth Edition. Copyright © 1989 by The Psychological Corporation.
Reproduced by permission. All rights reserved.

When displaying NCE results, many schools show results for grade levels over time as shown in Figure 5.5. One can make few comparisons within grade levels over time because the students are not the same. However, one usually can see the scores are fairly stable, unless something very different happens, such as using new teaching strategies or adding new teachers. Grades three and four illustrate the type of increases we would like to see every year.

NCE scores can be used for comparisons because they—

- have equal intervals;
- can be aggregated, disaggregated, and averaged;
- have a derived average of 50 and a standard deviation of 21.06;
- can be compared from one grade to another;
- can be used to calculate gain scores;
- match percentiles of 1 to 99;
- can be converted to percentiles after analysis.

Figure 5.5

EXAMPLE GRAPH: NCE RESULTS BY GRADE LEVEL

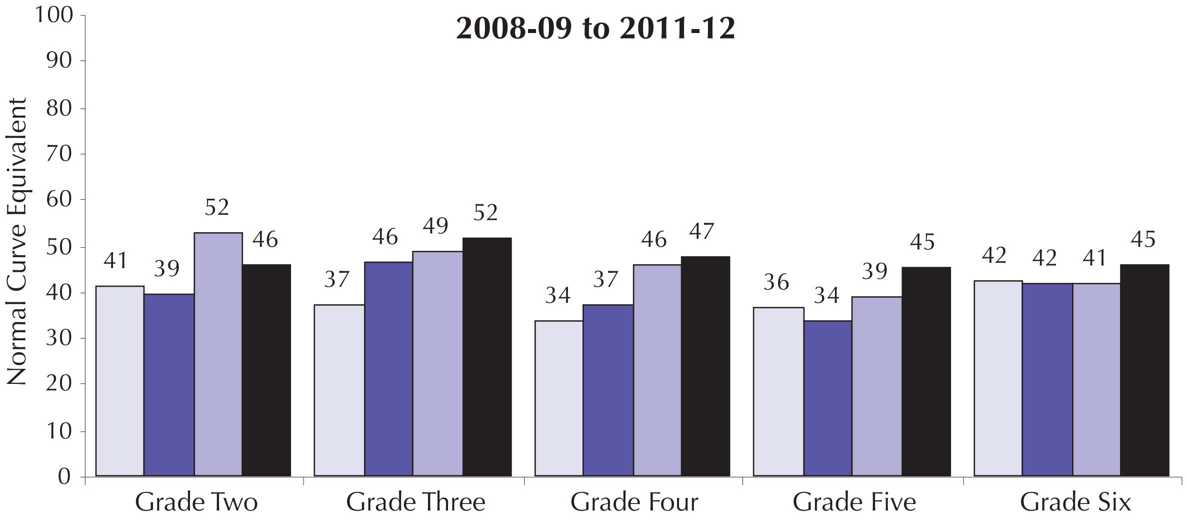
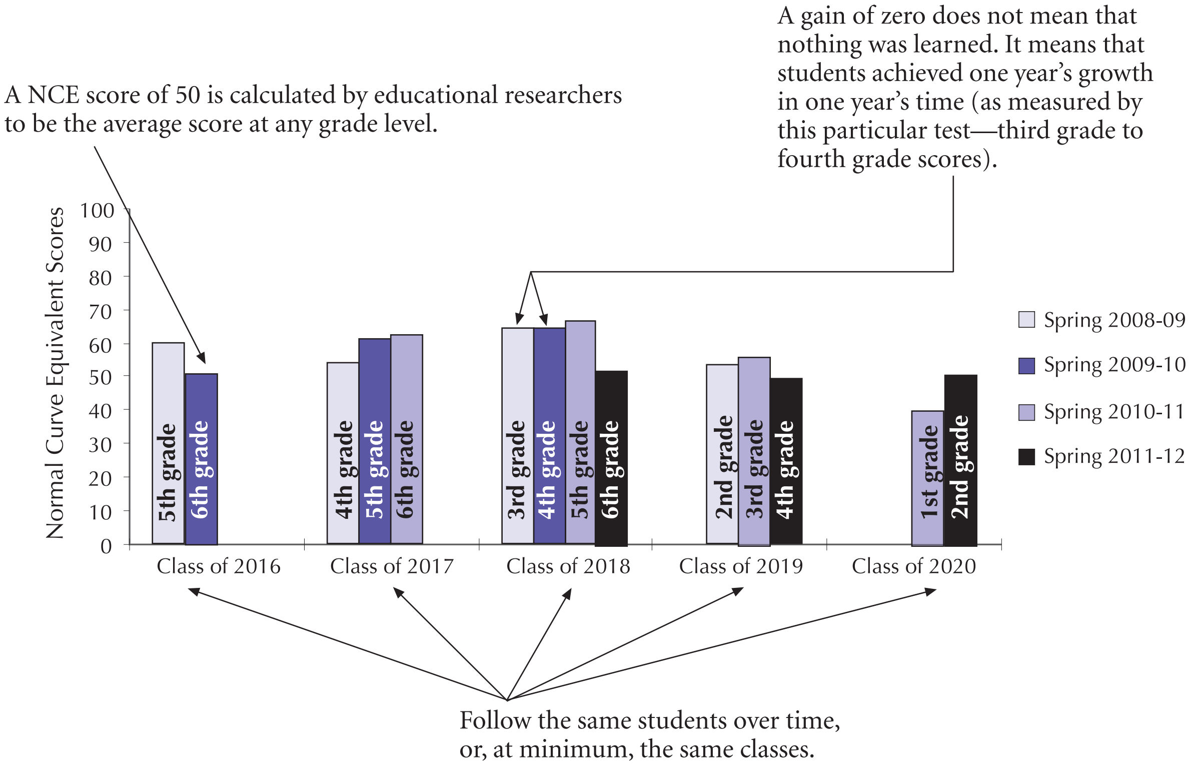


Figure 5.6 shows how to reorganize a grade level over time graph to follow the same group of students. This type of graph is called a "cohort graph." The cohorts can be matched (following the same students over time), or unmatched (following the same groups over time).

Figure 5.6
EXAMPLE GRAPH: MATCHED NCE MATH SCORES BY CLASS



A NCE score of 50 is calculated by educational researchers to be the average score at any grade level.

A gain of zero does not mean that nothing was learned. It means that students achieved one year's growth in one year's time (as measured by this particular test—third grade to fourth grade scores).

Follow the same students over time, or, at minimum, the same classes.

THE MEANING OF THE BLOOM QUOTE

> ***The normal curve is a distribution most appropriate to chance and random activity. Education is a purposeful activity and we seek to have students learn what we would teach. Therefore, if we are effective, the distribution of grades will be anything but a normal curve. In fact, a normal curve is evidence of our failure to teach.***
>
> **Benjamin Bloom**

The normal curve serves the purpose of showing us how our students score in relationship to norming groups.

The normal curve serves the purpose of showing us how our students score in relationship to norming groups. It says that when we give a test to many students, few will score very low, few will score very high, and most will score around the average score.

"*. . . A normal curve is evidence of our failure to teach.*" What this means is that as teachers, we have the challenge to instruct and to ensure that students achieve what we want them to know and be able to do. Our objective with teaching should be to make sure that the normal curve does not exist in our classroom. We want all students to learn the information.

Our objective with teaching should be to make sure that the normal curve does not exist in our classroom.

Standardized Test Score Terms

Standardized assessments have uniformity in content, administration, and scoring. Standardized test score terms are defined in Figure 5.7, along with a description of most effective uses and cautions for use.

Figure 5.7
STANDARDIZED TEST SCORE TERMS, THEIR MOST EFFECTIVE USES, AND CAUTIONS FOR THEIR USES

Score	Definition	Most Effective Uses	Cautions
Anticipated Achievement Scores	A student's anticipated achievement score is an estimate of the average score for students of similar ages, grade levels, and academic aptitude. It is an estimate of what we would expect the student to score on an achievement test.	Anticipated achievement scores can be used to see if a student is scoring "above" or "below" an expected score, indicating whether or not she/he is being challenged enough, or if her/his needs are being met.	It is easy to think of these scores as "IQ" scores. They are just achievement indicators on a standardized test.
Cognitive Abilities or Skills Index	The cognitive skills index is an age-dependent normalized standard score based on a student's performance on a cognitive skills test with a mean of 100 and standard deviation of 16. The score indicates a student's overall cognitive ability or academic aptitude relative to students of similar age, without regard to grade level.	Cognitive skills index scores can be used to see if a student is scoring "above" or "below" an expected score, indicating whether or not she/he is being challenged enough, or if her/his needs are being met.	It is easy to think of these scores as "IQ" scores. They are just achievement indicators on a standardized test.
Criterion-Referenced Tests	Tests that judge how well a test-taker does on an explicit objective relative to a predetermined performance level.	Tell us how well students are performing on specific criteria, goals, or standards.	CRTs test only what was taught, or planned to be taught. CRTs can not give a broad estimate of knowledge.
Deciles	Deciles divide a distribution into ten equal parts: 1–10; 11–20; 21–30; 31–40; 41–50; 51–60; 61–70; 71–80; 81–90; 91–99. Just about any scale can be used to show deciles.	Deciles allow schools to show how all students scored throughout the distribution. One would expect a school's distribution to resemble a normal curve. Watching the distribution move to the right, over time, could imply that all students in the distribution are making progress.	One must dig deeper to understand if all students and all groups of students are moving forward.
Diagnostic Tests	Diagnostic tests, usually standardized and normed, are given before instruction begins to help the instructor(s) understand student learning needs. Many different score types are used with diagnostic tests.	Help teachers know the nature of students' difficulties, but not the cause of the difficulty.	Make sure the diagnostic test is measuring what you want it to measure and that it can be compared to formative and summative assessments used.

Figure 5.7 *(Continued)*
STANDARDIZED TEST SCORE TERMS, THEIR MOST EFFECTIVE USES, AND CAUTIONS FOR THEIR USES

Score	Definition	Most Effective Uses	Cautions
Grade-Level Equivalents	Grade-level equivalents indicate the grade and month of the school year for which a given score is the actual or estimated average. Based on a ten-month school year, scores would be noted as 3.1 for third grade, first month, or 5.10 for fifth grade, tenth month.	Grade-level equivalents are most effectively used as a snapshot in time. Scores are comparable across subtests.	These scores should not be taken literally. If a third grader scored a 5.8 on a subtest, that does not mean that she/he should be doing fifth grade, eighth-month work. It only means that the student obtained the same score that one would expect average fifth-grade students in their eighth month of school to score if they took the third-grade test.
Latent-Trait Scale	A latent-trait scale is a scaled score obtained through one of several mathematical approaches collectively known as Latent-Trait Procedures or Item Response Theory. The particular numerical values used in the scale are arbitrary, but higher scores indicate more knowledgeable students or more difficult items.	Latent-trait scales have equal intervals allowing comparisons over time.	These are scores set up by testing professionals. Laypeople typically have difficulty understanding their meaning.
NCE (National or Local)	Normal Curve Equivalent (NCE) scores are standard scores with a mean of 50, a standard deviation of 21.06, and a range of 1 to 99. The term National would indicate that the norming group was national; local usually implies a state or district norming group.	NCEs have equal intervals so they can be used to study gains over time. The scores have the same meaning across subtests, grade levels, and years. A 50 is what one would expect in an average year's growth.	This score, just like all scores related to norm-referenced tests, cannot be taken literally. The score simply shows relative performance of a student group or of students to a norming group.
Norm-Referenced Tests	A norm-referenced test is a type of standardized test, assessment, or evaluation which yields an estimate of the position of the tested individual in a predefined population, with respect to the trait being measured.	Norm-referenced tests create meaning through comparing the test performance of a school, group, or individual with the performance of a norming group. The tests allow us to compare a student's skills to others in her/his age group.	Some norm-referenced score types are unequal interval scales, and should not be added, subtracted, or averaged.
Percent Passing	Percent passing is a calculated score implying the percentage of the student group meeting and exceeding some number, usually a cut score, proficiency/mastery level, or a standard.	With standards-based accountability, it is beneficial to know the percentage of the population meeting and exceeding a standard and to compare a year's percentages with the previous year(s) to understand progress being made.	This is a very simple statistic, and its interpretation should be simple as well. Total numbers (n=) of students included in the percentage must always be noted with the percentage to assist with the understanding.

Figure 5.7 *(Continued)*
STANDARDIZED TEST SCORE TERMS, THEIR MOST EFFECTIVE USES, AND CAUTIONS FOR THEIR USES

Score	Definition	Most Effective Uses	Cautions
Percentile / Percentile Rank (PR) (National or Local)	Percentile ranks indicate the percentage of students in a norm group (e.g., national or local) whose scores fall below a given score. The range is from 1 to 99. 50th percentile ranking would mean that 50 percent of the scores in the norming group fall below a specific score. The term National would indicate that the norming group was national; local usually implies a state or district norming group.	One-year comparison to the norming group. Schools can see the relative standing of a student or group in the same grade to the norm group who took the test at a comparable time.	Percentile rank is not a score to use over time to look for gains because of unequal intervals, unless the calculations are made with equal interval scores and then converted to percentile ranks. One cannot calculate averages using NPR because of the unequal intervals. Medians are the most appropriate statistic to use.
Proficiency / Performance Levels	Proficiency or performance levels describe student achievement results, related to ranges of scores on an assessment, usually as *below basic, basic, proficiency, advanced.*	Proficiency levels show the numbers and percentages of students scoring proficient and above, and the numbers and percentages of students scoring below proficiency -- with an indication of how far below – close/basic or not so close/below basic. Proficiency levels are good for displaying the distributions of results for a group and its subgroups.	Without the scores, proficiency levels do not provide an idea of where individual students scored within the proficiency ranges. Cannot calculate proficiency levels. Proficiency levels change by test.

Quartiles	There are three quartiles—Q1, Q2, Q3 — that divide a distribution into four equal groups:
	Q1=25th percentile
	Q2=50th percentile (Median)
	Q3=75th percentile
		Quartiles allow schools to see the distribution of scores for any grade level, for instance. Over time, schools trying to increase student achievement would want to monitor the distribution to ensure that all students are making progress.	With quartiles, one cannot tell if the scores are at the top of a quartile or the bottom. There could be "real" changes taking place within a quartile that would not be evident.

Raw Scores	Raw scores are the number of questions answered correctly on a test or subtest. A raw score is simply calculated by adding the number of questions answered correctly. The raw score is a person's observed score.	The raw score provides information about the number of questions answered correctly. To get a perspective on performance, raw scores must be used with the average score for the group and/or the total number of questions. Alone, it has no meaning.	Raw scores do not provide information related to other students taking the test or to other subtests. One needs to keep perspective by knowing the total number possible. Raw scores should never be used to make comparisons between performances on different tests unless other information about the characteristics of the tests are known and identical.

Figure 5.7 *(Continued)*

STANDARDIZED TEST SCORE TERMS, THEIR MOST EFFECTIVE USES, AND CAUTIONS FOR THEIR USES

Score	Definition	Most Effective Uses	Cautions
RIT Scale Scores	RIT scores, named for George Rasch who developed the theory of this type of measurement, are scaled scores that come from a series of tests created by the Northwest Evaluation Association (NWEA). The tests, which draw from an item bank, are aligned with local curriculum and state/local standards.	RIT scores provide ongoing measurement of curriculum standards and a way for students to see progress in their knowledge. The scores can also be shown as percentiles to know performance related to other students of similar ages and/or grades. You will most probably see gains each time a measurement is taken with a group of students.	RIT scores are great as long as the test was carefully designed to measure standards.
Scaled Scores	A scaled score is a mathematical transformation of a raw score.	The best uses of scaled scores are averages and averages calculated over time allowing for the study of change. These scores are good to use for calculations because of equal intervals. The scores can be applied across subtests on most tests. Scaled scores facilitate conversions to other score types.	Ranges vary, depending upon the test. Watch for the minimum and maximum values. It is sometimes hard for laypeople to create meaning from these scores. The normal curve is needed to interpret the results with respect to other scores and people.
Standard Scores	Standard score is a general term referring to scores that have been "transformed" for reasons of convenience, comparability, ease of interpretation, etc. z-scores and T-scores are standard scores.	The best uses of standard scores are averages and averages calculated over time, allowing for the study of change. These scores are good to use for calculations because of equal intervals. The scores can be applied across subtests on most tests. Scaled scores facilitate conversions to other score types.	Ranges vary, depending upon the test. Watch for the minimum and maximum values. It is sometimes hard for laypeople to create meaning from these scores. The normal curve is needed to interpret results with respect to other scores and people.
Standards-Based Assessments	Standards-based assessments measure students' progress toward mastering local, state, and/or national content standards.	The way standards-based assessments are analyzed depends upon the scales used. The most effective uses are in revealing the percentage of students achieving a standard.	One has to adhere to the cautions of whatever test or score type used. It is important to know how far from mastering the standard the students were when they did not meet the standard.

Figure 5.7 *(Continued)*
STANDARDIZED TEST SCORE TERMS, THEIR MOST EFFECTIVE USES, AND CAUTIONS FOR THEIR USES

Score	Definition	Most Effective Uses	Cautions
Stanines	Stanines are a nine-point standard score scale. Stanines divide the normal curve into nine equal points: 1 to 9.	Stanines, like quartiles, allow schools to see the distribution of scores for any grade level, for instance. Over time, schools trying to increase student achievement would want to monitor the distribution to ensure that all student scores are improving.	Often, the first three stanines are interpreted as "below average," the next three as "average," and the top three as "above average." This can be misleading. As with quartiles, one cannot tell if the scores are at the top of a stanine or the bottom. There could be "real" changes taking place within a stanine that would not be evident.
T-scores	A T-score is a standard score with a mean of 50 and a standard deviation of 10. T-scores are obtained by the following formula: $T=10z + 50$	The most effective uses of T-scores are averages and averages calculated over time. T-scores are good to use for calculations because of their equal intervals. T-scores can be applied across subtests on most tests because of the forced mean and standard deviation.	T-scores are rarely used because of the lack of understanding on the part of most test users.
z-scores	A z-score is a standard score with a mean of zero and a standard deviation of one. z-scores are obtained by the following formula: $z= \frac{\text{raw score } (x) - \text{mean}}{\text{standard deviation } (sd)}$	z-scores can tell one how many standard deviations a score is away from the mean. z-scores are most useful, perhaps, as the first step in computing other types of standard scores.	z-scores are rarely used by the lay public because of the difficulty in understanding the score.

MEASURING COMMON CORE STATE STANDARDS

With the adoption of the Common Core State Standards, most states in the United States will be changing the way they assess student achievement. Two consortia are tasked with creating student assessment systems aligned to a common core of academic content standards that:

- balance summative, interim, and formative testing through an integrated system of standards, curriculum, assessment, and instruction;
- effectively gauge college and career readiness; and
- support quick turnaround of results.

With the adoption of the Common Core State Standards, most states in the United States will be changing the way they assess student achievement.

As of the writing of this book, example assessments have not been released to allow us to describe how best to analyze and use the results. Watch for articles and addenda on the *Education for the Future* website: *http://eff.csuchico.edu.*

It is imperative that the appropriate analyses be used for the specific score type.

ANALYZING THE RESULTS, DESCRIPTIVELY

Descriptive statistics (i.e., mean, median, percent correct) can give schools very powerful information. It is imperative that the appropriate analyses be used for the specific score type. Figure 5.8 summarizes terms of analyses, their definitions, their most effective uses, and cautions for their uses in analyzing student achievement scores descriptively.

Descriptive statistics summarize the basic characteristics of a particular distribution, without making any inferences about population parameters. Graphing the information can also be considered descriptive.

Descriptive statistics summarize the basic characteristics of a particular distribution, without making any inferences about population parameters.

Figure 5.8

TERMS RELATED TO ANALYZING STUDENT ACHIEVEMENT RESULTS, DESCRIPTIVELY, THEIR MOST EFFECTIVE USES, AND CAUTIONS FOR THEIR USES

Term	Definition	Most Effective Uses	Cautions
Disaggregate	Disaggregation is breaking a total score into groups for purposes of seeing how subgroups performed. One disaggregates data to make sure all subgroups of students are learning.	Disaggregating student achievement scores by gender, ethnicity, backgrounds, etc., can show how different subgroups performed.	Disaggregations are for helping schools understand how to meet the needs of all students, not to say, "This group always does worse than the other group and always will." We must exercise caution in reporting disaggregations with small numbers in a subgroup.
Gain	Gain scores are the change or difference between two administrations of the same test. Gain scores are calculated by subtracting the previous score from the most recent score. One can have negative gains, which are actually losses.	One calculates gains to understand improvements in learning for groups of students and for individual students.	Gain scores should not be calculated using unequal interval scores, such as percentiles. The quality of gain score results is dependent upon the quality of the assessment instrument; the less reliable the assessment tool, the less meaningful the results. One needs to make sure the comparisons are appropriate, e.g., same students, same score types.
Item Analysis	Item analysis refers to the statistics surrounding students' performance on items on a test. Those statistics include item difficulty, discrimination, reliability, and distractor levels.	Item analysis is important for teachers to understand how students performed on each item, which response options received the most responses from students to reteach misconceived concepts.	Some of the statistics are hard for teachers to take the time to understand. It is important for teachers to know why students missed items.
Maximum	A maximum is the highest score achieved, or the highest possible score on a test.	Maximum possible scores and highest received scores are important for understanding the relative performance of any group or individual, especially when using scaled or standard scores.	A maximum can tell either the highest score possible or the highest score received by a test-taker. One needs to understand which maximum is being used in the analysis. It is best to use both.
Mean	A mean is the average score in a set of scores. One calculates the mean, or average, by summing all the scores and dividing by the total number of scores.	A mean can be calculated to provide an overall average for the group, and/or student, taking a specific test. One can use any equal interval score to get a mean.	Means should not be used with unequal interval scores, such as percentile ranks. Means are more sensitive to extreme results when the size of the group is small.
Median	A median is the score that splits a distribution in half: 50 percent of the scores fall above and 50 percent of the scores fall below the median. If the number of scores is odd, the median is the middle score. If the number of scores is even, one must add the two middle scores and divide by two to calculate the median.	Medians are the way to get a midpoint for scores with unequal intervals, such as percentile ranks. The median splits all scores into two equal parts. Medians are not sensitive to outliers, like means are.	Medians are relative. Medians are most effectively interpreted when reported with the possible and actual maximum and minimum.

Figure 5.8 *(Continued)*

TERMS RELATED TO ANALYZING STUDENT ACHIEVEMENT RESULTS, DESCRIPTIVELY, THEIR MOST EFFECTIVE USES, AND CAUTIONS FOR THEIR USES

Term	Definition	Most Effective Uses	Cautions
Minimum	A minimum is the lowest score achieved, or the lowest possible score on the test.	Minimum possible scores and lowest received scores are important for understanding the relative performance of any group or individual.	A minimum tells either the lowest score possible or the lowest score received by a test-taker. One needs to understand which minimum is being used. It is best to use both.
Mode	The mode is the score that occurs most frequently in a scoring distribution.	The mode basically tells which score or scores appear most often.	There may be more than one mode. The mode ignores other scores.
Percent Correct	Percent correct is a calculated score implying the percentage of students meeting and exceeding some number, usually a cut score, or a standard.	This calculated score can quickly tell educators how well the students are doing with respect to a specific set of items. It can also tell educators how many students need additional work to become proficient.	Percent correct is a calculated statistic, based on the number of items given. When the number of items given is small, the percent correct can be deceptively high or low.
Percent Proficient ***Percent Passing*** ***Percent Mastery***	Percent proficient, passing, or mastery represent the percentage of students who passed a particular test at a "proficient," "passing," or "mastery" level, as defined by the test creators or the test interpreters.	With standards-based accountability, it is beneficial to know the percentage of the population meeting and exceeding the standard and to compare a year's percentage with the previous year(s) to understand progress being made.	This is a very simple statistic, and its interpretation should be simple as well. Total numbers (N=) of students included in the percentage must always be noted with the percentage to assist in understanding the results. Ninety percent passing means something very different for 10 or 100 test-takers.
Range	Range is a measure of the spread between the lowest and the highest scores in a distribution. Calculate the range of scores by subtracting the lowest score from the highest score. Range is often described as end points also, such as the range of percentile ranks is 1 and 99.	Ranges tell us the width of the distribution of scores. Educators working on continuous improvement will want to watch the range, of actual scores, decrease over time.	If there are outliers present, the range can give a misleading impression of dispersion.
Raw Scores	Raw scores refer to the number of questions answered correctly on a test or subtest. A raw score is simply calculated by adding the number of questions answered correctly. The raw score is a person's observed score.	The raw score provides information only about the number of questions answered correctly. To get a perspective on performance, raw scores must be used with the average score for the group and the total number of questions. Alone, raw scores have little meaning.	Raw scores do not provide information related to other students taking the test or to other subtests or scores. One needs to keep perspective by knowing the total number possible. Raw scores should never be used to make comparisons between performances on different tests unless other information about the characteristics of the tests are known and identical.

Figure 5.8 *(Continued)*

TERMS RELATED TO ANALYZING STUDENT ACHIEVEMENT RESULTS, DESCRIPTIVELY, THEIR MOST EFFECTIVE USES, AND CAUTIONS FOR THEIR USES

Term	Definition	Most Effective Uses	Cautions
Relationships	Relationships refer to looking at two or more sets of analyses to understand what they mean to each other without using extensive statistical techniques.	Descriptive statistics lend themselves to looking at the relationships of different analyses to each other; for instance, student learning results disaggregated by ethnicity, compared to student questionnaire results disaggregated by ethnicity.	This type of analysis is general and the results should be considered general as well. This is not a "correlation."
Standard Deviation	The standard deviation is a measure of variability in a set of scores. The standard deviation indicates how far away scores are from the mean. The standard deviation is the square root of the variance. Unlike the variance, the standard deviation is stated in the original units of the variable. Approximately 68 percent of the scores in a normal distribution lie between plus one and minus one standard deviation of the mean. The more scores cluster around the mean, the smaller the variance.	Tells us about the variability of scores. Standard deviations indicate how spread-out the scores are without looking at the entire distribution. A low standard deviation would indicate that the scores of a group are close together. A high standard deviation would imply that the range of scores is wide.	Often this is a confusing statistic for laypeople to understand. There are more descriptive ways to describe and show the variability of student scores, such as with a decile graph. Standard deviations only make sense with scores that are distributed normally.
Triangulation	Triangulation is a term used for combining three or more measures to get a more complete picture of student achievement.	If students are to be retained based on standards proficiency, educators must have more than one way of knowing if the students are proficient or not. Some students perform well on standardized measures and not on other measures, while others do not do well with standardized measures. Triangulation allows students to display what they know on three different measures.	It is sometimes very complicated to combine different measures to understand proficiency. When proficiency standards change, triangulation calculations will need to be revised. Therefore, all the calculations must be documented so they can be recalculated when necessary.

ANALYZING THE RESULTS, INFERENTIALLY

Many school administrators and teachers have taken statistics courses that taught them it is important to have control groups, experimental designs, and to test for significant differences. These terms fall in the category of inferential statistics. Inferential statistics are concerned with measuring a sample from a population, and then making estimates, or inferences, about the population from which the sample was taken. Inferential statistics help generalize the results of data analysis when one is not using the entire population in the analysis.

Descriptive analyses provide helpful and useful information and can be understood by a majority of people. When using the entire school population in your analyses, there is no need to generalize to a larger population—you have the whole population. There is little need for inferential statistics when doing basic data analysis work.

Inferential statistical methods, such as analyses of variance, correlations, and regression analyses are complex and require someone who knows statistics to meet the conditions and verify the conformity to the assumptions (postulates) of the analyses. Since there are times when a statistician is available to perform inferential statistics, some of the terms the statistician might use with tests include those listed in Figure 5.9.

Inferential statistics are concerned with measuring a sample from a population, and then making estimates, or inferences, about the population from which the sample was taken.

Inferential statistical methods, such as analyses of variance, correlations, and regression analyses are complex and require someone who knows statistics to meet the conditions and verify the conformity to the assumptions (postulates) of the analyses.

Figure 5.9

TERMS RELATED TO ANALYZING STUDENT ACHIEVEMENT RESULTS, INFERENTIALLY, THEIR MOST EFFECTIVE USES, AND CAUTIONS FOR THEIR USES

Term	Definition	Most Effective Uses	Cautions
Analysis of Variance (ANOVA)	Analysis of variance is a general term applied to the study of differences in the application of approaches, as opposed to the relationship of different levels of approaches to the result. With ANOVAs, we are testing the differences of the means of at least two different distributions.	ANOVAs can be used to determine if there is a difference in student achievement scores between one school and another, keeping all other variables equal. It cannot tell you what the differences are, per se, but one can compute confidence intervals to estimate these differences.	Very seldom are the conditions available to study differences in education in this manner. Too many complex variables get in the way, and ethics may be involved. There are well-defined procedures for conducting ANOVAs to which we must adhere.
Correlation Analyses	Correlation is a statistical analysis that helps one understand the relationship of scores in one distribution to scores in another distribution. Correlations show magnitude and direction. Magnitude indicates the degree of the relationship. Correlation coefficients have a range of -1.0 to +1.0. A correlation of around zero would indicate little relationship. Correlations of .8 and higher, or -.8 and lower would indicate a strong relationship. When the high scores in one distribution are also high in the comparing distribution, the direction is positive. When the high scores in one distribution are related to the low scores in the other distribution, the result is a negative correlational direction.	Correlations can be used to understand the relationship of different variables to each other, e.g., attendance and performance on a standardized test; .40 to .70 are considered moderate correlations. Above .70 is considered to be high correlations.	It is wise to plot the scores to understand if the relationship is linear or not. One could misinterpret results if the scores are not linear. Pearson correlation coefficient requires linear relationships. Also, a few outliers could skew the results and oppositely skewed distributions can limit how high a Pearson coefficient can be. Also, one must remember that correlation does not suggest causation.
Regression Analyses	Regression analysis results in an equation that describes the nature of the relationship between variables. Simple regression predicts an object's value on a response variable when given its value on one predictor variable. Multiple regression predicts an object's value on a response variable when given its value on each of several predictor variables. Correlation tells you strength and direction of relationship. Regression goes one step further and allows you to predict.	A regression equation can be used to predict student achievement results, for example. Regression can determine if there is a relationship between two or more variables (such as attendance and student background) and the nature of those relationships. This analysis helps predict and prevent student failure, and predict and ensure student successes.	One needs to truly understand the statistical assumptions that need to be in place in order to perform a regression analysis. This is not an analysis to perform through trial and error.

Figure 5.9 *(Continued)*
TERMS RELATED TO ANALYZING STUDENT ACHIEVEMENT RESULTS, INFERENTIALLY, THEIR MOST EFFECTIVE USES, AND CAUTIONS FOR THEIR USES

Term	Definition	Most Effective Uses	Cautions
Control Groups	During an experiment, the control group is studied the same as the experimental group, except that it does not receive the treatment of interest.	Control groups serve as a baseline in making comparisons with treatment groups. Control groups are necessary when the general effectiveness of a treatment is unknown.	It may not be ethical to hold back from students some method of learning that we believe would be useful.
Experimental Design	Experimental design is the detailed planning of an experiment, made beforehand, to ensure that the data collected are appropriate and obtained in a way that will lead to an objective analysis, with valid inferences.	Experimental designs can maximize the amount of information gained, given the amount of effort expended.	Sometimes it takes statistical expertise to establish an experimental design properly.
Experimental Group	An experimental group is a group of individuals who are part of a clinical study or experiment who are exposed to the treatments of the experiment, while another group, the control group, is not.	Experimental groups are most effectively used in education when win-win situations are created. In other words, the control group will not be at a disadvantaged because they did not receive the treatment.	We do not always need experimental groups to test theories of what is working and what is not working. We can study results with all students.
Tests of Significance	Tests of significance use samples to test claims about population parameters.	Tests of significance can estimate a population parameter, with a certain amount of confidence, from a sample.	Often lay people do not know what *statistically significant* really means.

MEASUREMENT ERROR

All measurement, by definition, contains some error. However, that error can be measured.

We often hear people talk about sample sizes being too small to make any conclusions. Typically, what this means is that the larger the number of students in a sample, the more confidence we have that the score is an accurate reflection of that sample group's abilities. We also want to know if the increases showing up in a testing program in any year are because of "true" increases in learning. In order to understand a "true" gain resulting from an instructional program, we construct confidence bands or intervals that give each score a range as opposed to a single number. This range provides flexibility in understanding what the score would be if there were "errors" in the test. The table in Figure 5.10 gives the calculated standard error of measure for normal NCEs.

The larger the number of students in a sample, the more confidence we have that the score is an accurate reflection of that sample group's abilities.

Figure 5.10
GIVE AND TAKE TABLE

Number of Students	Error (NCEs)	Number of Students	Error (NCEs)
1	10.1	20	2.3
2	7.1	25	2.0
3	5.8	30	1.8
4	5.1	35	1.7
5	4.5	40	1.6
6	4.1	45	1.5
7	3.8	50	1.4
8	3.6	75	1.2
9	3.4	100	1.0
10	3.2	200	0.7
15	2.6	300	0.6

Stanford Achievement Test: Eighth Edition. Copyright © 1989 by The Psychological Corporation. Reproduced by permission. All rights reserved.

To construct a confidence band, consider that a group of five students had an average score of 40 on a reading test. We need to look at Figure 5.10, the *Give and Take Table,* to see that the error of measurement associated with the five students in this group is 4.5—the smaller the number of students, the more likely it is that the average represents a chance occurrence or error. The confidence band would then be formed from 35.5 to 44.5 (i.e., 40 - 4.5 = 35.5; 40 + 4.5 = 44.5). The interpretation: We can feel confident (68% of the time—one standard deviation) that the average true score of these students would be somewhere between 35.5 and 44.5. We can feel 95 percent confident that the average true score would be between 31 and 49 (double the size of the band, or two standard deviations, 2 x 4.5 = 9) (i.e., 40 - 9 = 31; 40+ 9 = 49).

VALIDITY AND RELIABILITY OF ASSESSMENTS

Other important terms associated with testing are defined below. Please see the *References and Resources* section at the end of this book for further information about these terms.

VALIDITY

The validity of a test or assessment refers to whether it provides the type of information desired. Validity can be enhanced by asking appropriate questions that get to what you want to know.

The validity of a test or assessment refers to whether it provides the type of information desired.

Ways to document and demonstrate validity include the following:

- Content validity relates to the appropriateness of the items with respect to the content, instruction, or the curriculum being measured.
- Predictive validity refers to a test's ability to predict future performance in the area that the instrument is measuring.
- Face validity relates to the appearance that the test (or the items on the test) measures what it claims to measure.
- Construct validity refers to the degree to which the test actually measures the particular construct (trait or aptitude) in question.
- Concurrent validity refers to the scores on a test being related to currently existing measures of the same content or behavior.
- Differential validity refers to the degree to which a test does not have a bias that would favor a particular sub-group of individuals (boys versus girls).
- Consequential validity refers to the wanted and unwanted implications resulting from the use of a test.

RELIABILITY

The reliability of a test or assessment relates to the consistency with which knowledge is measured. Reliability tells us that if students were to take the test more than once, they would get the same (or nearly the same) score.

The reliability of a test or assessment relates to the consistency with which knowledge is measured.

Reliability is impacted, among other things, by—

- the range of individual differences (a test administered to a heterogeneous group will increase the reliability of the results);
- the length of the test in terms of number of items (reliability increases with length);
- the time limit (if the test is too long and students cannot complete all the questions, the answers at the end, the ones not completed, will show high correlation because the scores are usually 0. That will artificially show an undue increase in reliability);
- the range of item difficulty (very easy tests or very difficult tests will have a lower reliability);
- the unidimensionality (all items should measure the same dimension, trait, or construct); and

- the consistency of the testing environment and the testing material (booklets) among test sites and in time. Reliability increases when sources of errors are limited and controlled.

A test may be reliable without being valid, but it cannot be valid without being reliable.

A test may be reliable without being valid, but it cannot be valid without being reliable. As an example, a tape measure is an extremely "reliable" tool when measuring the length of a wall. However, it would be a non-"valid" measuring tool when preparing the mix for a cake. Also, a kitchen scale is a "valid" tool to measure the weight, but would be non-reliable and non-"valid" at measuring an adult's body weight.

HOW MUCH TIME DOES IT TAKE?

Most schools have summative student learning data available through their state or provincial offices of education. The Data Team might want to spend a couple of hours reorganizing the data for the data profile.

Organizing multiple measures of student learning data will take a little longer, starting with identifying what is being used in all grade levels and subject areas.

REFLECTION QUESTIONS

1. What are student learning data?
2. Why are student learning data important for continuous school improvement?
3. How is your school assessing student learning, and why? Are these approaches appropriate?
4. Who should know the student learning data of a school?

APPLICATION OPPORTUNITIES

1. Use the student learning data inventory (Appendix B3) to list how your school assesses student learning. Note gaps in assessments and uses of assessments. Make sure your assessments are appropriate for the intended uses and information.
2. Graph your schoolwide student achievement data, disaggregated in the ways mentioned in the chapter and as shown in Appendix F.